

Дәріс 5. Ықтималдық және гипотезалар.

Оқытушы: Сарсембаева Т.С.

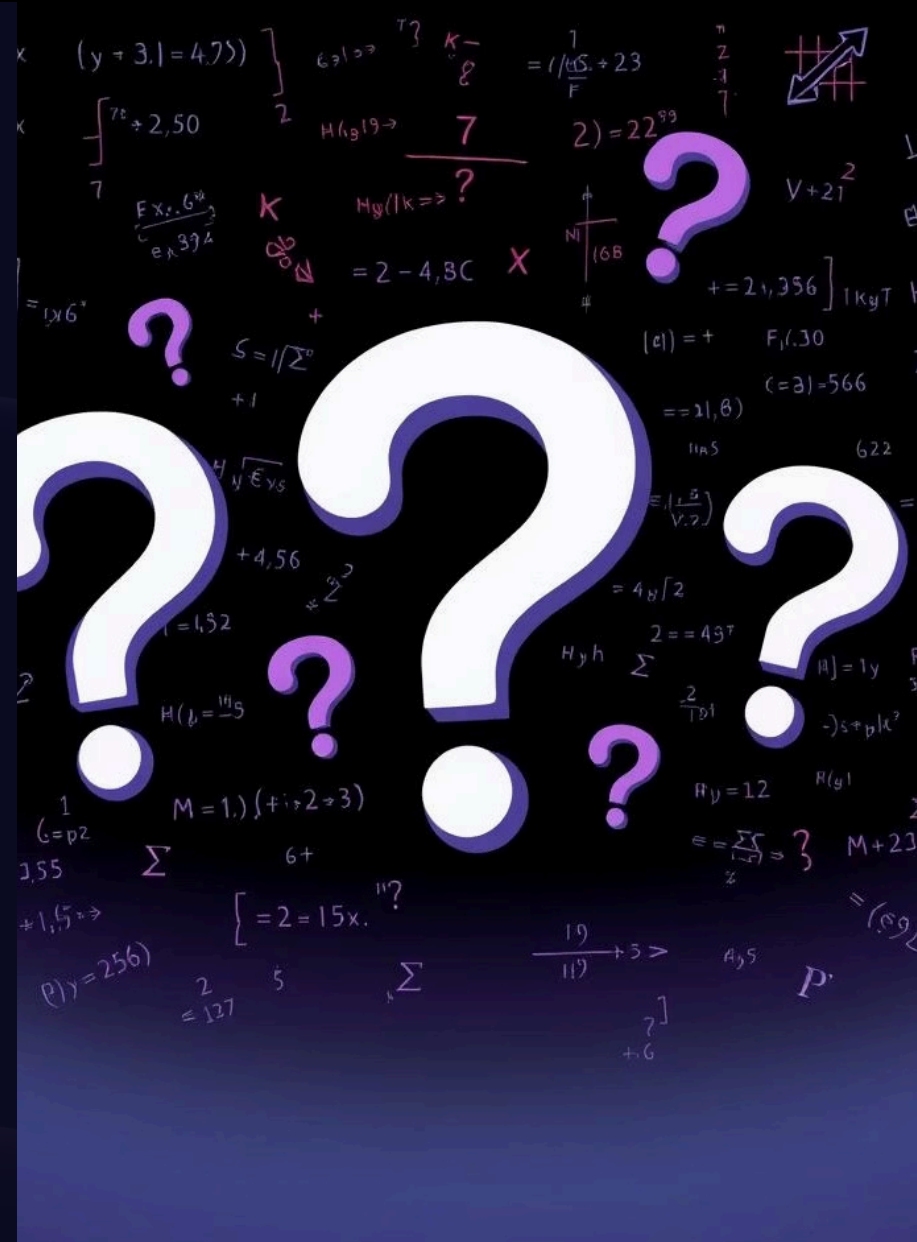


Дәрістің мақсаты:

Ықтималдық теориясының деректерді талдаудағы және белгісіздік жағдайында шешім қабылдаудағы рөлімен таныстыру. Одан кейін, ықтималдықтың негізгі ұғымдарын, соның ішінде тәуелді және тәуелсіз оқиғаларды, шартты ықтималдықты және Байес теоремасын меңгерту көзделеді. Сондай-ақ, дәріс математикалық статистиканың маңызды тұжырымдамалары – Қалыпты үлестіру мен Орталық шектеу теоремасын түсіндіруге бағытталған. Соңында, статистикалық қорытынды жасаудың негіздерін, соның ішінде гипотезаны тексеру (H_0/H_1), p -мәндер, сенімділік аралықтары және A/B сынақтарын оқыту мақсат етіледі.

Негізгі сұрақтар:

1. Шартты ықтималдық дегеніміз не және ол тәуелді оқиғалармен қалай байланысты?
2. Байес теоремасы қалай жұмыс істейді және оның негізгі компоненттері қандай?
3. Қалыпты үлестіру дегеніміз не және Орталық шектеу теоремасының маңызы неде?
4. Гипотезаны тексеру процесі қалай жүреді?
5. Р-мән (p-value) дегеніміз не және ол шешім қабылдауға қалай әсер етеді?
6. Сенімділік аралығы (Confidence Interval) және A/B сынағы дегеніміз не?





Кіріспе

Математика мен статистика деректерді талдаудың, машиналық оқытудың және жасанды интеллектпен жұмыс істеудің негізі болып табылады.

Деректер талдаушылары мен деректер ғылымдарының мамандары көбінесе ықтималдық теориясына сүйенеді, өйткені бұл оларға белгісіздік жағдайында дұрыс шешім қабылдауға мүмкіндік береді.

Мысалы, жобаның бастапқы кезеңінде, нақты деректер болмаған кезде, ең тиімді гипотезаны таңдау үшін Ықтималдықтар мен болжамдар қолданылады.

Ықтималдық

Ықтималдықтар теориясы — бұл модельдеу мен криптография үшін кездейсоқ сандарды құруға, алгоритмдер мен мәліметтер құрылымын жақсартуға, сонымен қатар машиналық оқытудың нақты модельдерін жасауға көмектесетін маңызды құрал.

$$P(A) = \sum_i P(A|B_i)P(B_i)$$

Бүгінгі дәрісте ықтималдылықты белгілі бір оқиғаға байланысты белгісіздікті сандық бағалау әдісі ретінде қарастырамыз. Осылай сіз мұны қарапайым мысалдан техникалық анықтамалар мен математикалық дәл тұжырымдарға терең үңілмей-ақ түсіне аласыз.

Мысалы, текшені қалай айналдыру керектігін елестетіп көрейік.

Ықтималдық кеңістігі - бұл барлық мүмкін болатын негізгі нәтижелердің жиынтығы. Текше жағдайында мұндай нәтижелер алтыға тең: {1,2,3,4,5,6}

Және бұл нәтижелердің кез келген бөлігі кездейсоқ оқиға болып саналады.

Мысалы: "бір саны түсті" - бұл бір элементтен тұратын оқиға {1};

"жұп санның түсі" - бұл үш элементтен тұратын оқиға {2,4,6}.

Яғни, ықтималдық әрбір осындай оқиғаның пайда болу ықтималдығын сандық түрде сипаттауға мүмкіндік береді.

Өзара тәуелділік және тәуелсіздік

Ықтималдық теориясында оқиғалар бір-бірімен тәуелді немесе тәуелсіз болуы мүмкін. Бұл ұғымдар оқиғалардың өзара байланысын сипаттайды, яғни: бір оқиғаның пайда болуы немесе болмауы екінші оқиғаның ықтималдығына әсер етеді ме, жоқ па — соны анықтайды.

Тәуелсіз оқиғалар (Independent events):

Егер бір оқиғаны жүзеге асыру екіншісінің ықтималдығын өзгертпесе, екі оқиға тәуелсіз деп аталады.

Басқаша айтқанда, бір оқиғаның нәтижесі екіншісіне ешқандай әсер етпейді.

$$P(A \cap B) = P(A) \times P(B)$$

$P(A)$ - А оқиғасының ықтималдығы, $P(B)$ - В оқиғасының ықтималдығы, $P(A \cap B)$ - екі оқиғаның бір уақытта пайда болу ықтималдығы.

Тәуелді оқиғалар (Dependent events):

Егер бір оқиғаның пайда болуы екіншісінің ықтималдығын өзгертсе, онда олар тәуелді оқиғалар деп аталады.

$$P(A \cap B) = P(A) \times P(B|A)$$

$P(B|A)$ — Егер А оқиғасы орын алса, В оқиғасының ықтималдығы.

Шартты ықтималдық (Conditional probability)

Шартты ықтималдық - бір оқиға В болғаны белгілі болған тағы бір оқиға А ықтималдықты білдіретін көрсеткіш. Ол "В жағдайында А ықтималдығы" оқылады және белгіленеді:

$$P(A|B)$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ егер } P(B) > 0.$$

Яғни, А және В ықтималдығы, екеуі де $P(A \cap B)$ В ықтималдығына бөледі.

Байес теоремасы (Bayes' Theorem)

Байес теоремасы - ықтималдық теориясының ең маңызды және кеңінен қолданылатын құралдарының бірі болып табылады.

Ол ағылшын ғалымы, діни қызметкер және математик Томас Байестің (Томас Байес, 1702-1761) есімімен аталды.

1763 жылы ғалым қайтыс болғаннан кейін оның жұмысы “An Essay towards solving a Problem in the Doctrine of Chances” деген атпен жарық көрді.

Бұл жұмыста Байес жаңа ақпарат немесе деректер пайда болған кезде сенімдер мен ықтималдықтарды қалай жаңарту керектігін көрсететін принципті ұсынды.

Байес теоремасының мәні оқиға туралы жаңартылған сенімділікті (posterior probability) есептеу үшін алдын ала сенімділікті (prior probability) және жаңа деректерді (evidence) біріктіру болып табылады.

Басқаша айтқанда, біз бұрыннан белгілі ықтималдықтар негізінде жаңа деректерді алған кезде өз көзқарасымызды сандық түрде түзетеміз.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

$P(A)$ - априорлық ықтималдық (prior), яғни оқиға A -ның бастапқы ықтималдығы;

$P(B|A)$ - шартты ықтималдық, A болған жағдайда B -ның ықтималдығы;

$P(B)$ - барлық мүмкін жағдайларда B ықтималдығы (нормализация коэффициенті);

$P(A|B)$ - апостериорлық ықтималдық (posterior), яғни, B дерегі анықтағаннан кейін A -ның ықтималдығы.

Қалыпты үлестіру (Normal distribution) және Орталық шектеу теоремасы (Central Limit Theorem)

Қалыпты үлестіру - кездейсоқ шамалардың таралуын сипаттау үшін ең көп қолданылатын модель болып табылады.

Оның кең таралуының себебі - нақты өмірдегі кез-келген өлшем нәтижесінде қателіктердің пайда болуы. Бұл қателіктерге көбінесе бірнеше факторлардың қосындысы әсер етеді (мысалы, өлшеу құралының дәлдігі, адамның қателігі, дөңгелектеу және т.б.).

Сондықтан, егер айнымалы әр түрлі кішігірім әсерлердің қосындысынан тұрса, оның таралуы қалыпты заңға жуықтайды.

Орталық шектеу теоремасы - ықтималдық теориясы мен математикалық статистиканың маңызды заңдарының бірі.

Бұл теорема кездейсоқ шамалардың үлкен санының қосындысы (немесе орташа мәні) бастапқы шамалардың таралуы тең болса да, қалыпты үлестірімге жақын екенін көрсетеді.

$$X_i, i = 1, 2, \dots, n$$

Айталық, - бір-бірінен тәуелсіз және бірдей таралған кездейсоқ шамалар, олардың орташа мәні μ , дисперсиясы σ^2 (шекті мәнге ие, яғни $\sigma^2 < \infty$).

$$E(X) = \sum_{i=1}^n \mu_i \quad \text{and} \quad V(X) = \sum_{i=1}^n \sigma_i^2$$

Нормальды үлестірімнің ықтималдық тығыздығы келесі формуламен беріледі:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

μ — орташа мән (mean), σ — стандартты ауытқу (standard deviation), σ^2 — дисперсия (variance).

Гипотеза және қорытынды (Hypothesis and Conclusion)

Гипотеза - бұл белгілі бір құбылысты, мәліметтер арасындағы байланысты немесе Параметрлер мәндерін болжауға, түсіндіруге бағытталған ғылыми пайымдау.

Яғни, гипотеза - зерттеу жүргізер алдында зерттеуші ұсынған болжам немесе болжам, содан кейін ол нақты деректермен тексеріледі.

Ғылыми тұрғыдан алғанда, гипотеза-бұл эксперименттер мен статистикалық әдістер арқылы дәлелдеуді қажет ететін белгілі бір құбылыстың себебін немесе заңдылығын уақытша түсіндіретін пікір.

Түрлері:

Нөлдік гипотеза (Null hypothesis, H_0): бұл бастапқы, стандартты болжам. Ол зерттелетін деректер арасында ешқандай айырмашылық немесе байланыс жоқ дейді.

H_0 : "айырмашылық жоқ", "өзгеріс жоқ", "параметр белгілі мәнге тең"

Мысал:

H_0 : студенттердің орташа баллы = 70 балл.

Баламалы гипотеза (Alternative hypothesis, H_1 немесе H_a)

Бұл айырмашылық немесе әсер бар деп болжайтын нөлдік гипотезаның қарама-қайшы нұсқасы.

H_1 : "айырмашылық бар", "өзгеріс бар", "параметр белгілі мәнге тең емес"

Мысал: H_1 : студенттердің орташа баллы $\neq$ 70 балл.

Гипотезаны тексеру

Гипотезаны тексеру-бұл деректерге негізделген нөлдік гипотезаны қабылдау немесе жоққа шығару процесі.

Қадамдар:

1. Гипотезаларды құру: H_0 және H_1 анықтамасы.
2. Маңыздылық деңгейін таңдау: Әдетте $\alpha=0,05$ немесе 5%. Бұл теріске шығару ықтималдығы кател 0 қатемен.
3. Сынақ статистикасын есептеу:(мысалы: t тесті, z тесті, β_2 тесті және т. б.)
4. Р-мәнін табу (p-value):Бұл нәтиженің кездейсоқ шығу ықтималдығы.
5. Шешім қабылдау: Егер $p \leq \alpha \rightarrow$ теріске 0 жоққа шығарылады. Егер $p > \alpha \rightarrow 0$ болса H_0 қабылданады.

Р-мән (p-value) — бұл нөлдік гипотеза (H_0) дұрыс деп есептеген жағдайда, алынған немесе одан да “шектен тыс” нәтижені кездейсоқ түрде алу ықтималдығы.

Басқаша айтқанда: р-мән – “осындай нәтижені тек кездейсоқтықтың салдарынан алу мүмкін бе?” деген сұрақтың сандық жауабы.

Егер p мәні өте аз болса, бұл нәтижені қарапайым кездейсоқтықпен түсіндіру қиын, сондықтан біз нөлдік гипотеза шындыққа жанаспайды деген қорытындыға келеміз.

Егер p мәні үлкен болса, нәтиже кездейсоқ ауытқудың салдары болуы мүмкін, сондықтан нөлдік гипотезаны жоққа шығаруға негіз жоқ.

Маңыздылық деңгейі (α) - зерттеушінің қателік жіберетін шегі.

Сенімділік аралықтары (Confidence Intervals)

Сенімділік аралықтары(интервалы) - статистикалық есептеулер негізінде табылған параметрдің шын мәні жатуы мүмкін диапазон.

Яғни, біз нақты мәнді білмейміз, бірақ белгілі бір ықтималдықпен (мысалы, 95%) біз оны осы аралықта деп сенімді түрде айта аламыз.

Басқаша айтқанда: Сенімділік интервалы - бұл параметрдің мүмкін мәндерінің интервалдық бағасы (орташа, үлес, айырмашылық және т.б.).

Мысал қарастырайық: бізде 100 студенттің орташа баллы $\bar{x}=70$, стандартты ауытқу $s=10$

Сенімділік деңгейі-95%.

Сенімділік аралығы келесідей есептеледі:

$$CI = \bar{x} \pm z \frac{s}{\sqrt{n}}$$

$\bar{x}$ - таңдамалық орташа мән, s — стандартты ауытқу, n — таңдама көлемі, $z_{\alpha/2}$ — стандартты нормаль таралудан алынған критикалық мән (95% үшін 1.96).

Біз 95% сенімдіміз, яғни нақты орташа мән (популяцияның орташа мәні) 68,04 пен 71,96 арасында. $CI=[68.04;71.96]$

A/B сынақтары (A/B Testing)

A / B сынағы-бұл екі немесе одан да көп нұсқаларды салыстыру әдісі (мысалы, веб-парақ, жарнама, дизайн, өнім нұсқалары) және қайсысы тиімді екенін статистикалық анықтау.

Мақсаты: пайдаланушылардың нақты әрекеттеріне сүйене отырып, қай опция жақсы нәтиже беретіні туралы дәлелді табу.

Мысал ретінде сайтта екі батырма бар делік: А нұсқасы: жасыл түсті, В нұсқасы: қызыл түсті. Біз білгіміз келеді — қай түсті батырма көбірек басылады?

Гипотеза бойынша қарастыратын болсақ нөлдік гипотеза (H_0): А және В нұсқалары еш айырмашылығы жоқ десе, баламалы гипотеза (H_1): В нұсқасы жақсы нәтиже береді (мысалы, жоғары конверсия) деп болжамдады.

Ендігі қай нұсқа дұрыс екенін анықтау үшін біз тест өткіземіз:

1. Біріншіден мақсатты метриканы анықтап аламыз;
2. Екіншіден қолданушыларды екі топқа бөлеміз(A/B);
3. Үшіншіден деректерді жинаймыз;
4. Төртіншіден статистикалық талдау жасаймыз;
5. Соңынан барып қорытынды жасаймыз.

Қорытынды (Conclusion)

Қорытынды-гипотезаны тексеру нәтижесінде алынған шешім. Бұл ғылыми тұжырым немесе деректер негізінде қабылданған шешім.

Қорытынды гипотезаны дәлелдемейді немесе жоққа шығармайды, тек деректерге сәйкес оны қолдайтынын көрсетеді.

Дәрісті меңгеруге арналған бақылау сұрақтары:

1. Шартты ықтималдық дегеніміз не және оның формуласы қандай?
2. Байес теоремасының негізгі мақсаты неде? (Априорлық және апостериорлық ықтималдықтарды түсіндіріңіз).
3. Орталық шектеу теоремасы неліктен маңызды болып саналады?
4. Нөлдік (H_0) және баламалы (H_1) гипотезалардың айырмашылығы неде? Мысал келтіріңіз.
5. Р-мәні (p-value) 0.03-ке тең болса (және маңыздылық деңгейі $\alpha = 0.05$ болса), сіз нөлдік гипотезаға (H_0) қатысты қандай шешім қабылдайсыз?
6. А/В тестілеу дегеніміз не және ол гипотезаны тексерумен қалай байланысты?

Әдебиеттері тізімі:

- Грас, Дж. (2020). Data Science. Наука о данных с нуля.
https://www.ursmu.ru/userfiles/media/default/sveden/education/1.2.1_iiimo-21_mm_kaf.inf.pdf
- Yandex Practicum блогы. (2021). Теорема Байеса для Data Science. <https://practicum.yandex.ru/blog/teorema-bajesa-dlya-data-science/>
- Google Drive құжаты. (2022). Data Science негіздері. <https://share.google/jeo6osrS81LY4lhqP>
- West Georgia университеті. (2020). Confidence Intervals Notes.
https://www.westga.edu/academics/research/vrc/assets/docs/confidence_intervals_notes.pdf